American Journal of Epidemiology

Vol. 192, No. 5
https://doi.org/10.1093/aje/kwad030
Advance Access publication:
February 14, 2023

Practice of Epidemiology

Novel Methods for Leveraging Large Cohort Studies for Qualitative and Mixed-Methods Research

Katie Truc Nhat H. Nguyen, Jennifer J. Stuart, Aarushi H. Shah, Iris A. Becene, Madeline G. West, Jane Berrill, Bizu Gelaye, Christina P. C. Borba, and Janet W. Rich-Edwards*

* Correspondence to Dr. Janet W. Rich-Edwards, Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1 Brigham Circle, Boston, MA 02115 (e-mail: JR33@partners.org).

Initially submitted December 15, 2022; accepted for publication February 1, 2023.

Qualitative research methods, while rising in popularity, are still a relatively underutilized tool in public health research. Usually reserved for small samples, qualitative research techniques have the potential to enhance insights gained from large questionnaires and cohort studies, both deepening the interpretation of quantitative data and generating novel hypotheses that might otherwise be missed by standard approaches; this is especially true where exposures and outcomes are new, understudied, or rapidly changing, as in a pandemic. However, methods for the conduct of qualitative research within large samples are underdeveloped. Here, we describe a novel method of applying qualitative research methods to free-text comments collected in a large epidemiologic questionnaire. Specifically, this method includes: 1) a hierarchical system of coding through content analysis; 2) a qualitative data management application; and 3) an adaptation of Cohen's κ and percent agreement statistics for use by a team of coders, applying multiple codes per record from a large codebook. The methods outlined in this paper may help direct future applications of qualitative and mixed methods within large cohort studies.

COVID-19; free-text comments; interrater reliability; large questionnaires; mixed methods; qualitative data management; qualitative research; research methods

Abbreviations: COVID-19, coronavirus disease 2019; IRR, interrater reliability; PPE, personal protective equipment.

> *There are more things in heaven and earth, Horatio, than are dreamt of in your [questionnaire].*
>
> Adapted from *Hamlet* (1)

Epidemiologists are experts at measuring quantitative traits across populations using assessment instruments that require study participants to choose from a set of categorical response options (2). Questionnaires enable researchers to efficiently accumulate vast amounts of quantitative data from large samples. However, such "checkbox epidemiology" can frustrate study participants and investigators alike. To minimize participant burden and maximize participation, questionnaires are usually kept as brief as possible (3). Investigators often "leave on the table" questions of interest they were unable to squeeze into a questionnaire. When questionnaires require participants to distill complex life events into checkbox responses that fail to capture the totality of their experiences, misclassification and missing data can ensue. Another limitation of relying exclusively on quantitative measures is the lost opportunity for unexpected insights from participants: Investigators tend to query about only the domains for which there already exist questionnaire instruments (4). In unusual and fast-changing times, such as during a pandemic, it may be difficult to anticipate all the exposures and outcomes experienced by study populations that will inform public health and medicine. Further, when circumstances are evolving rapidly, an instrument designed at one moment in time may prove outdated by the time it is fielded.

In contrast, qualitative research includes techniques for collecting, organizing, and interpreting nonnumerical data from in-depth interviews, focus groups, conversations, written text, and visual formats (e.g., photography, drawings). The goal of qualitative research is to understand phenomena by studying participants' own interpretations of their experiences (5). Whereas quantitative research aims to address

prespecified problems through deductive reasoning, qualitative research relies on inductive probing to capture the rich depth and diversity of experiences, utilizing broad prompts to afford individuals freedom in their responses (6). Qualitative methodology creates thematic categories that emerge directly from the data and analyzes their relationships (7). This is especially informative when investigators want to understand the context of complex, multifactorial public health problems; to pursue research in populations or content where survey instruments have not been developed or validated; or to examine emerging diseases or understudied populations and exposure-outcome relationships (8). However, because the purpose is to understand deep contextual information, qualitative approaches are traditionally modest in sample size ($n << 200$) (9) and prioritize homogeneity over generalizability (10), rendering them less attractive to epidemiologists.

While the superiority of quantitative or qualitative research is widely debated, both approaches have strengths and weaknesses (4, 11). Recent studies have leveraged the strengths of both methods through mixed-methods research (12). However, sample size can be a challenge for mixed-methods studies because, traditionally, quantitative research demands large samples to detect statistically significant differences and generalize findings, while qualitative research samples focus on depth and rarely exceed 200 participants (9, 13).

There is an opportunity to apply qualitative research methods to large populations by taking advantage of free-text comment boxes on questionnaires. Current questionnaires often include comment boxes for logistical purposes, such as detecting address changes or giving participants a place to vent about unsatisfactory items. Such comment boxes may also offer an overlooked opportunity to collect qualitative data. Especially with the use of prompts, free-text comments might be mined to enrich a study's findings, help inform future research questions, and elevate the voices of study participants as part of the research process. Qualitative research may provide contextual information to complement quantitative findings. While investigators may skim participant comments to monitor complaints or mine a few anecdotes, this approach misses the chance to rigorously explore themes raised by participants through application of systematic qualitative research methods that reduce bias, improve insight, and increase reproducibility. To our knowledge, no other studies have analyzed open-ended qualitative data from very large survey samples ($n > 1,000$), and methods have yet to be developed for this context.

Our objective was to develop a methodological approach for analyzing open-ended qualitative data collected from a large participant sample. In this paper, we describe a novel process which was developed and used to apply qualitative research methods to a series of coronavirus disease 2019 (COVID-19) questionnaires returned by 58,614 participants in 3 US longitudinal studies (Nurses' Health Study II, Nurses' Health Study 3, and the Growing Up Today Study), of whom 32,947 (56%) contributed free-text comments. (See Web Appendix 1, available at https://doi.org/10.1093/aje/kwad030, for more details on application of this method to our specific study, including participant flow and survey response (Web Appendix 1, Web Figure 1) and characteristics of respondents and nonrespondents (Web Appendix 1, Web Table 1).)

METHODS

Design and positioning of open text boxes

Electronic questionnaires can include any number of free-text comment boxes with prompts of varying specificity. For example, the baseline questionnaire of our COVID-19 survey included 4 comment text boxes: 2 unprompted boxes labeled "Comments" following questions about COVID-19 symptoms and diet-related questions; a specific prompted box about personal protective equipment (PPE) that was tied to a study aim ("Please include any information about your use of improvised, non-standard PPE"); and a more general prompted box at the end of the questionnaire ("We are interested in learning more about your experiences during this pandemic. Please add anything else you would like to tell us here."). The length of the free-text comments should not be restricted; in reviewing tens of thousands of comments, very few were longer than a short paragraph. The placement of the comment boxes after specific items is likely to invite comments related to those questions; for example, we received many comments about immune-boosting supplements in the unprompted box that followed the dietary assessment.

Overview: codebook development and qualitative data analysis

Qualitative content analysis (14, 15) can be used to interpret the meaning of participants' free-text comments, using codes (topic labels) derived and assigned by trained individuals reviewing the text (coders) based on their consensus interpretation of the text. This involves an iterative process performed by the coding team consisting of coding, memoing, calculating interrater reliability (IRR), and testing and revising a codebook dictionary. This process continues until data saturation is reached (i.e., when no new information or insights emerge from reading additional records) (16). Coders purposively sample participants across groups (e.g., age, sex, race/ethnicity, geography, or exposures of interest (e.g., occupational or pregnancy status)) within the cohort to ensure saturation is based on a representative sample of the data (17).

Coding: open, axial, and selective

Three phases of coding are performed: open, axial, and selective. The process starts with open coding, in which a team of coders independently reviews free-text comments from the same set of records (e.g., $n = 200$) to identify preliminary codes. Using content analysis, each coder employs an inductive approach to open-code these free-text comments, individually identifying and defining their own set of preliminary codes (18).

After open coding, axial coding occurs as coders convene to compare their individual preliminary codes, explore

relational patterns qualitatively detected among the codes while reviewing the text responses, and organize them under broader categories, called "parent codes." Our initial codebook sought to broadly capture all emergent themes relevant to the pandemic, resulting in a codebook dictionary with over 150 codes. However, coders in later projects with more focused aims (e.g., experiences of pregnancy among healthcare workers during the pandemic) created their own codebooks with a limited number of focused codes. These "parent codes" and the "codes" nested under them formulate the first version of the codebook dictionary. The coders establish definitions for each code, articulate inclusion and exclusion criteria (when to use and not use each code), and identify an example of correct application. Table 1 depicts an example of several codes that were organized under the parent code of "Child Care and Concern" in our COVID-19 study.

Once the first version of the codebook dictionary is created, coders proceed to selective coding. In this third stage, coders use the codebook dictionary to assign codes to both previously reviewed and newly sampled records. While the primary focus of selective coding is not to create new codes, coders make note of new themes that emerge to discuss with the coding team after each round of coding, which may result in revisions to the codebook dictionary. After each round of revision, more records are sampled and coded, and previously reviewed records are reviewed again to apply any changes in each iteration of the codebook dictionary. For efficiency, if the round of revision only modified existing codes (e.g., redefined the definitions of existing codes and inclusion/exclusion criteria, renamed codes, merged codes) without adding new codes, then coders rereviewed only records that had been assigned modified codes. In cases where distinct new codes were added, all previously reviewed records were rereviewed. IRR statistics are calculated in the selective coding phase to evaluate each iteration of the codebook dictionary. Specifics of the IRR calculation and codebook revision process are detailed below.

Calculating interrater reliability: the IRR application

The quality of the codebook dictionary is monitored by frequent checks of IRR during the selective coding phase. Low IRR results prompt revision of the codebook dictionary. Two measures of IRR are commonly used:

1. Percent agreement ($\Pr(a)$), which is calculated by dividing the number of agreements (records to which coders applied the same code) by the total number of agreements plus disagreements (records to which coders did not apply the same code), and
2. Cohen's kappa (κ) (19, 20), which accounts for percent chance agreement ($\Pr(e)$) and is calculated as $\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$.

For efficiency, a large set of records requires a team of coders. Cohen's κ, conventionally used to measure IRR, is limited in that it is designed to compare 1) only 2 coders and 2) the assignment of only 1 code per record (20). Fleiss (21) introduced an adaption that permitted multiple coders. However, Fleiss' κ requires that each record be assigned the same number of codes. We wished to allow coders the flexibility to assign multiple and varying numbers of codes per record, so as not to constrain the depth and completeness of the codebook. Recently, Krippendorff's α has been proposed as an alternative IRR statistic to accommodate various numbers of codes assigned per unit of text and numbers of coders (22); however, it is not yet technically possible to calculate this statistic within a large data set (Dr. Richard Craggs, University of Leicester (Leicester, United Kingdom), personal communication, 2021). Thus, we adapted Cohen's κ, following the method used by NVivo (QSR International, Burlington, Massachusetts), the industry standard for qualitative data analysis, to derive average κ values and $\Pr(a)$ across multiple coders and codes (23). The specific calculations and their automation are fully described in Web Appendix 2, which depicts the calculations of IRR between any 2 coders (Web Appendix 2, Web Figure 2) and across coders (Web Appendix 2, Web Figure 3). The spreadsheet for calculating IRR statistics, as well as written and video guides to its use, are included in Web Appendices 3 and 4 and Web Video 1. Briefly, we sought to generate the following κ statistics and their 95% confidence intervals.

- Within each pair of coders:
  - κ for each code (within-pair, each code (WPEC)), κ_{WPEC}
  - Average κ across all codes (within-pair, all codes (WPAC)), κ_{WPAC}
- Across all coders:
  - Average κ for each code (across pairs, each code (APEC)), κ_{APEC}
  - Grand average (GA) κ across all codes and coders, κ_{GA}

$\Pr(a)$ was calculated in a similar manner.

For both IRR statistics, we scored agreement for codes at the participant level—that is, across all free-text comments provided by each participant. (We use "record" to refer to all available qualitative responses provided by a given participant across all free-text comment boxes.) An alternative would be to score agreement across each free-text comment box separately. The IRR statistics were calculated as the arithmetic means between pairs of coders, across all coders for each code, and as a grand average across all coders and codes. These IRR statistics have varying functions in developing the codebook, training new coders, and reporting final codebook reliability.

Testing and revising the codebook dictionary

The iterative process of developing the codebook dictionary in the selective coding phase follows the cycle depicted in Figure 1.

First, for each iteration of the dictionary, coders apply any new codes to both the previously coded records and a new set of records. IRR is tested through "double-coding" of new records by multiple coders. For example, each of the 5 coders in our study was assigned 200 records to code, plus 20

Table 1. Excerpt From the Codebook Dictionary[a]

Code	Definition	When to Use (Inclusion Criteria)	When Not to Use (Exclusion Criteria)	Example
Homeschooling challenges: parental responsibility	Participant describes challenges that arise from the increased responsibility of having to homeschool children resulting from school closures	Use when participant discusses any difficulties that arise from their added responsibility of having to educate and homeschool children	Do not use when participant discusses feeling stressed and concerned about homeschooling when it pertains to the quality of materials and its effect on children	"Working from home while caring for two small children and trying to homeschool has been very stressful, for my spouse and I."
Homeschooling challenges: quality	Participant describes feeling stressed about the lower-quality educational resources for children	Use when participant discusses feeling stressed and/or concerned about the lower quality of homeschooling	Do not use when participant discusses feeling stressed about homeschooling as an added responsibility	"E-learning has made this very very stressful for young children. Young children do not connect well over web-based platforms such as Zoom."
Child response stress	Participant describes experiencing stress related to their child(ren)'s adverse response to the pandemic	Use when participant describes challenges and stress that arise from their child(ren) dealing/coping with the new pandemic lifestyle	Do not use when participant discusses feeling stressed about added responsibilities pertaining to children, including education and general child care	"The hardest part has been managing the anxiety of my family . . . not specifically COVID-related but as a result of the isolation. [My] 7-year-old son needs socializing and is lonely, [my] 13-year-old daughter has anxiety that has worsened, and now she has daily physical complaints that she thinks are life-threatening. [My] husband's anxiety is also high and he's becoming withdrawn. It's been hard to help them all from spiraling downward. I am a nurse but not in direct patient care anymore, and I have severe feelings of guilt because I did not go to NYC to assist because I chose to stay with my family. If I were single or my children were grown, I would have gone."
Child impact/well-being concern	Participant describes feeling concerned about their child(ren)'s well-being	Use when participant expresses concern about how the pandemic is adversely affecting the physical, emotional, mental, and psychosocial well-being of their children	Do not use when participant discusses concern about children's education	"Need a gym. Because of a bad hip, it's hard to trust trails with kids. Need my physio and massage and to physically see specialists. I forget things when it's over the phone and kids are running around. Kids at home and hard to teach and force to play. One has autism traits but not diagnosed but have behavior care plan at school so we get no help or respite. Oldest is social and needs friends around. My job was to be from home and is on and off because of hip and limitations with exposure to COVID, as I have health conditions to where I shouldn't be around it and it's limited because of poor organization and lack of supplies for home use."

Table continues

Table 1. Continued

Code	Definition	When to Use (Inclusion Criteria)	When Not to Use (Exclusion Criteria)	Example
Child care and parenting stress	Participant describes stress in response to general parenting and child-care responsibilities	Use when participant attributes current stress or frustration to their general responsibilities in parenting and caring for their children	Do not use if participant attributes current stress to other domestic factors not directly related to child care, or to factors involving specifically their child's education	"Supporting teens during this time is trying. We have 1 child and feel it's more difficult to support her and ourselves during this unknown time. It seems that we all have our days, when I'm up someone else is down."

Abbreviations: COVID, coronavirus disease 2019; NYC, New York City.

[a] Individual codes nested under the parent code "Child Care and Concern" and the corresponding definitions for each code, inclusion and exclusion criteria, and an example of correct application.

records from each of the other 4 coders' sets of 200 assigned records (for a total of 280 records reviewed by each coder). Thus, a total of 1,000 records were reviewed by the coding team, and 100 of these 1,000 records were coded by all 5 coders to test the IRR.

Second, after coding, IRR is calculated. Investigators should decide upon a target/minimum IRR to guide when the codebook requires further revision (if the IRR falls below that target) or is adequate to proceed with coding of new records (if the IRR meets or exceeds that target). One such threshold might be $\kappa_{GA} \geq 0.6$; a traditional Cohen's κ value greater than or equal to 0.6 signifies at least a moderate level of agreement (20). If the prespecified threshold is met, coders proceed to review new records in batches of 100–300 per coder with periodic IRR checks of a subset of those records (e.g., 50–100) to ensure both continued agreement among coders and that no new codes emerge from new records. An IRR check is also initiated when a new coder is added to the team or a previous coder returns from a hiatus. If κ_{GA} falls below the threshold, coders should review codes for which κ_{APEC} was especially low (in our case, less than 0.4), evaluating the clarity of the code definition and the ease of recalling the code's definition based on its name (by focusing on codes with $\kappa_{APEC} < 0.4$, we were able to efficiently increase κ_{GA} to $\geq$0.6). At this stage, coders might discuss whether the infrequent use of some codes reflects the particular sample of records or whether the code would be unlikely to be applied in the larger set of records. If the latter, coders may merge it with another code or remove it from the codebook dictionary entirely. If the unused or infrequently used code is retained, coders should reevaluate it at the next IRR check.

Third, if changes are proposed, the coders revise the codebook dictionary to add or subtract codes, reorganize groupings, combine existing codes, or revise the definitions of existing codes. The cycle begins again with the (re)reviewing of records with the revised codebook until κ_{GA} reaches the criterion threshold.

This iterative process results in multiple versions of the codebook dictionary. By design, data saturation cannot be attained until the codebook dictionary is finalized. Once the coders reach data saturation, all themes have been identified and captured in the codebook dictionary and no further records are required to be sampled and coded. However, investigators may continue to code records to increase representation of their sample; we chose to code at least 5% of the total records with free-text comments, sometimes oversampling particular groups to improve our contextual knowledge of those groups.

Quantitative analysis usually compares participants who were in/eligible or who did/did not return a questionnaire or respond to a particular item. The sampling frame of an epidemiologic study affords the opportunity to compare the characteristics of participants who did and did not contribute comments, or whose comments were or were not selected for coding; this is not typical of qualitative studies but can be leveraged to further understand the data and the context of comments analyzed (e.g., see Web Table 1 in Web Appendix 1).

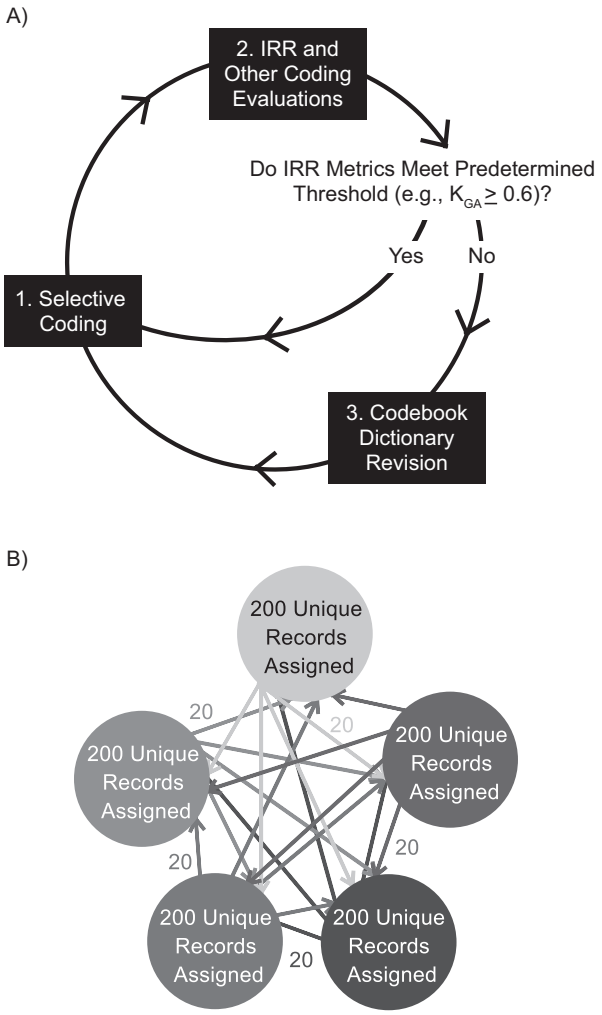

Figure 1. Schematic overview (A) of the iterative test and revision process used to develop the codebook dictionary using interrater reliability (IRR) metrics. Stage 1: selective coding of records with codes outlined in the codebook dictionary. "Double-coding" of records for IRR checks takes place. For example (B), each coder codes 280 records, the coding team reviews 1,000 records in total, and 100 of the 1,000 records are coded by all coders for the IRR check. Stage 2: an IRR check is performed. Coders discuss the coding process and propose new codes or changes to the codebook dictionary based on codes with low agreement. Stage 3: revision to accommodate changes discussed at the previous stage. κ_{GA} represents the grand average (GA) κ value across all codes and coders.

Developing the codebook application

To store, organize, and assign codes from the codebook dictionary to records and to quantify data at different levels of codes, we developed a new qualitative data management application using Microsoft Excel (Microsoft Corporation, Redmond, Washington), which we named the Codebook Application. A spreadsheet template for the Codebook Application and detailed information about its use are included in Web Appendices 5–7; a video guide to the codebook is included in Web Video 1. Commercial software that could be used for this process includes NVivo (24). However, reported drawbacks of NVivo include cost and time-consuming learning processes due to the overwhelming amount of functions (25). Unlike NVivo, our application is free to use and does not require coders to have proficient knowledge of any software (including Excel) to load thousands of participant records (including both quantitative data and free-text comment boxes) into the workbook, assign codes, and calculate IRR. The main skills required for our Excel-based Codebook Application are the ability to organize codes and parent codes in a table, select labels from a drop-down box, and copy and paste cells from one worksheet to another. In addition, while NVivo and similar software allows concurrent multiuser access so that multiple coders can code the same records at the same time, this collaboration depends on Internet connectivity (26). The Codebook Application integrates with our IRR Application, which, as previously discussed, innovates upon existing IRR methods (i.e., Cohen's κ, Fleiss' κ, Krippendorff's α) to accommodate various numbers of codes assigned per unit of text, multiple coders, and large data sets.

The Codebook Application allows coders to import participant records (identification and comment fields) into a spreadsheet and search for codes to assign to those records from drop-down menus that enable easy location of codes (in Web Appendix 6, Web Figure 4 shows a sample codebook entry; Web Figure 5 gives an example of parent and child codes; and Web Figures 6 and 7 depict the linkage of worksheets within the Codebook Application). Linked spreadsheets calculate IRR statistics, enabling coders to measure the quality of the codebook and identify inconsistently applied codes (Web Appendix 2).

Memoing: a practice of self-reflexivity

While analyzing comments, the coding team members record their thoughts and interpretations of the data through the process of memos. Memos, like a journal entry, include reflections on the analysis process, questions about ambiguous data, ideas about the codebook, and interpretations of and connections formed between larger themes revealed in the data. Memos aid in the process of developing consensus in the application of codes and code definitions between coders, as they enable each coder to articulate individual thoughts and prompt the group to examine multiple perspectives and proposed theories.

Qualitative analysis and manuscript preparation

Once coding is complete, investigators can perform purely qualitative or mixed-methods analyses. The selection of manuscript topics may be driven by high-frequency codes (reflecting their salience) or by novel insights gleaned by the coders during the analysis process.

With a large codebook, investigators need to select the most relevant codes to include in their analysis. For example, although we created a broad, general codebook dictionary covering all pandemic-related topics, for an article regarding health-care workers' use of PPE we chose to work with

5 codes under the parent code "PPE," 2 codes under the parent code "Virus Spread Concern," and 1 code under the parent code "Work Stressors." Typically, relevant codes are presented in a table in the publication with their definitions.

Codes can be used in many ways, including: to count the frequency of codes; to identify patterns across codes; and to gain contextualized information through illustrative quotations for inclusion in the manuscript. For example, the "PPE policy" code was applied to 6% of active health-care workers' records. We then examined the distribution of the code according to the quantitative domains captured in the questionnaire, including whether the participant had treated patients with COVID-19, whether the participant had adequate PPE access, and the COVID-19 mortality rate of the census region in which the participant lived. Finally, the following representative quotation was identified in a record assigned the "PPE policy" code:

> Our hospital administrators told us 'per the CDC' we didn't need N95s unless doing an aerosolizing procedure. They didn't routinely provide them at first. Five nurses got sick from this one patient. Our charge nurse was fired for speaking up about PPE.

The baseline questionnaire included a prompted comment box about PPE, so the many responses about PPE were expected. Other responses, particularly those from the prompt "Tell us about your experiences during the pandemic," were more surprising. Several insights detected through coding of early questionnaires prompted us to incorporate new quantitative questions in subsequent questionnaires regarding gratitude, furloughs, parenting/work conflicts, and discrimination against health-care workers.

The pluripotent nature of cohorts and open coding can yield many topics and treatments, ranging from purely qualitative approaches to mixed methods and largely quantitative analyses illustrated by quotations.

DISCUSSION

The application of qualitative research methods to participants' free-text comments allows participant perspectives to expand the breadth and depth of inquiry. The themes that emerge are probably broader than the restricted topics covered by most questionnaires; they inform the interpretation of quantitative data, permitting a triangulation across different types of data. Qualitative themes may suggest new hypotheses and prompt future data collection to test these hypotheses. Advantages of applying this approach in large studies over small, focused samples typical of qualitative research are the pluripotent comparisons and diverse perspectives enabled by the sheer number of respondents; for example, contrasts can be made according to factors such as demographic characteristics, occupational status, parental status, geographic region, and local rates of disease. Depending on the data set, it may be possible to explore the experiences of several intersectional marginalized identities. In this way, applying qualitative methods to a large study sample, as described here, can be analogous to conducting multiple qualitative studies in one, increasing efficiency and breadth of insights gained from a single participant population and data source (e.g., questionnaire). One limitation of the large sample-based qualitative research approach we propose here is its inability to probe participants' responses in real time. Traditional qualitative research projects with small-to-modest sample sizes can achieve greater depth of insight by responding to participant cues during interviews and focus groups to probe topics/themes that arise—something that is impossible in the context of a large survey.

There is an inherent tension between breadth and depth; the application of qualitative methods to large surveys carries both the best and the worst of its parent disciplines. In a 1993 editorial, Britten and Fisher noted, "There is some truth in the quip that quantitative methods are reliable but not valid and that qualitative methods are valid but not reliable" (27, p. 271). While qualitative research employs IRR to improve reproducibility, there are ongoing philosophical and practical debates about the purpose, choice, limitations, and interpretation of IRR statistics; these considerations are discussed in Web Appendix 8. The philosophical objection to IRR statistics is that they preference consensus above discovery; this debate also occurs regarding the handling and value of outlier data in quantitative sciences. Too much agreement among coders can reflect a lack of diversity of viewpoints and limit the generalizability of the codebook. Disagreements between coders may be as valuable as their consistency (28). There are also practical exceptions to IRR statistics. For example, critics of IRR statistics note their vulnerability to factors such as the length of the text segment, the number of codes within a codebook, the frequency with which codes are applied, and asymmetrical application of codes between coders (29). Thus, the importance of IRR may depend on the aims or context of a study. Ultimately, the best use of IRR in qualitative research may lie in the internal process of improving the codebook and training new coders.

Consensus and discovery are both involved in the research approach proposed here, but their contributions and value vary by stage. During the process of developing the codebook dictionary, discovery and consensus run in parallel, driving the iterative nature of the codebook dictionary revision process. Through discovery, new codes and identification of patterns are proposed by individual coders. By consensus, these discoveries are refined and incorporated into the codebook dictionary. Consensus ensures that all coders agree on the definition and proper application of the codes. Whereas quantitative researchers often identify and remove outliers for analysis, qualitative researchers commonly seek out and give voice to the apparent outliers. Information gained by probing discordant responses can provide valuable insights and ensure that the findings reported from qualitative or mixed-methods research accurately reflect the nuances of the lived experiences of participants.

The work detailed here adapted and extended traditional qualitative research techniques, typically used on modest-sized samples (9), for application to large questionnaires with prompted or unprompted free-text comment boxes. To our knowledge, this approach is novel in large epidemiologic studies. Adding a qualitative research component to large surveys may be especially useful in research endeavors

at the "edge" of our knowledge: Novel situations (e.g., a pandemic), emerging new diseases (e.g., "long COVID" or the new *Diagnostic and Statistical Manual of Mental Disorders* diagnosis of "prolonged grief disorder" (30)), cases where investigators realize that existing instruments map imperfectly onto complex phenomena, and the "lived experience" of people may be hard to capture in a 10-item survey. Qualitative research allows for unexpected findings and adaptation. From data collection to codebook development to IRR tests, these methods can facilitate the application of qualitative methods within large-scale population questionnaires to stimulate new breadth and depth of discovery, beyond what can be achieved with quantitative methods alone.

ACKNOWLEDGMENTS

Author affiliations: Department of Chemistry and Chemical Biology, Harvard College, Harvard University, Cambridge, Massachusetts, United States (Katie Truc Nhat H. Nguyen); Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States (Katie Truc Nhat H. Nguyen, Jennifer J. Stuart, Jane Berrill, Janet W. Rich-Edwards); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States (Jennifer J. Stuart, Bizu Gelaye, Janet W. Rich-Edwards); Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, New York, New York, United States (Aarushi H. Shah); NYU School of Global Public Health, New York University, New York, New York, United States (Madeline G. West); Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States (Jane Berrill); Chester M. Pierce Division of Global Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States (Bizu Gelaye); and Department of Psychiatry, School of Medicine, Boston University, Boston, Massachusetts, United States (Christina P. C. Borba).

C.P.C.B. and J.W.R.-E. are co–senior authors.

This work was supported by the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention (award 75D30120P08670); the National Institutes of Health (award T23HD049339-15); and the Radcliffe Research Partnership.

Our public websites (https://nurseshealthstudy.org/, https://www.nhs3.org/, and https://gutsweb.org/) include brief descriptions of the Nurses' Health Studies and Growing Up Today Study cohorts, all questionnaires, and a description of resource-sharing procedures. An automated online form requests that investigators applying for data briefly describe their study's hypothesis and aims, the variables needed, etc. Requests are presented to the cohort investigator meetings every other week, and replies are provided within 2 weeks. In general, if the cohorts have the data/resources required for a proposal, data-sharing is approved.

The views expressed in this article are those of the authors and do not reflect those of the National Institute for Occupational Safety and Health, the National Institutes of Health, or the Radcliffe Research Partnership.

Conflict of interest: none declared.

REFERENCES

1. Shakespeare WT. *The Tragedy of Hamlet, Prince of Denmark*. In: Wright WA, ed. *The Complete Works of William Shakespeare: The Cambridge Edition Text*. Garden City, NY: Doubleday, Doran & Company, Inc.; 1936:743.
2. Safdar N, Abbo LM, Knobloch MJ, et al. Research methods in healthcare epidemiology: survey and qualitative research. *Infect Control Hosp Epidemiol.* 2016;37(11):1272–1277.
3. Fan W, Yan Z. Factors affecting response rates of the web survey: a systematic review. *Comput Human Behav.* 2010; 26(2):132–139.
4. Lakshman M, Sinha L, Biswas M, et al. Quantitative vs qualitative research methods. *Indian J Pediatr.* 2000;67(5): 369–377.
5. Malterud K. Qualitative research: standards, challenges, and guidelines. *Lancet*. 2001;358(9280):483–488.
6. Allen M. *The SAGE Encyclopedia of Communication Research Methods*. Thousand Oaks, CA: SAGE Publications, Inc.; 2017.
7. Charmaz K. 'Discovering' chronic illness: using grounded theory. *Soc Sci Med.* 1990;30(11):1161–1172.
8. Haghani M, Bliemer MCJ. Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: scientometric comparisons across SARS, MERS and 2019-nCoV literature. *Scientometrics.* 2020; 125(3):2695–2726.
9. Emmel N. Sample size. In: *Sampling and Choosing Cases in Qualitative Research: A Realist Approach*. London, United Kingdom: SAGE Publications Ltd.; 2013:137–156.
10. Carminati L. Generalizability in qualitative research: a tale of two traditions. *Qual Health Res.* 2018;28(13):2094–2101.
11. Greene JC, Caracelli VJ, Graham WF. Toward a conceptual framework for mixed-method evaluation designs. *Educ Eval Policy Anal.* 1989;11(3):255–274.
12. Wisdom JP, Cavaleri MA, Onwuegbuzie AJ, et al. Methodological reporting in qualitative, quantitative, and mixed methods health services research articles. *Health Serv Res.* 2012;47(2):721–745.
13. Collins KMT, Onwuegbuzie AJ, Jiao QG. A mixed methods investigation of mixed methods sampling designs in social and health science research. *J Mixed Methods Res.* 2007;1(3): 267–294.
14. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res.* 2005;15(9):1277–1288.
15. Auerbach C, Silverstein L. *Qualitative Data: An Introduction to Coding and Analysis*. New York, NY: New York University Press; 2003.
16. Faulkner SL, Trotter SP. Data saturation. In: Matthes J, Davis C, Potter R, eds. *The International Encyclopedia of Communication Research Methods*. Newcastle Upon Tyne, United Kingdom: SAGE Publications Ltd.; 2017:1–2.

17. Glaser B, Strauss A. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine Publishing Company; 1967.
18. Gale NK, Heath G, Cameron E, et al. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol.* 2013; 13:117.
19. Miles MB, Huberman AM. *Qualitative Data Analysis. An Expanded Sourcebook*. 2nd ed. Thousand Oaks, CA: SAGE Publications, Inc.; 1994.
20. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012, 22;(3):276–282.
21. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378–382.
22. Krippendorff K, Craggs R. The reliability of multi-valued coding of data. *Commun Methods Meas.* 2016;10(4): 181–198.
23. QSR International. Run a coding comparison query. http://help-nv11.qsrinternational.com/desktop/procedures/run_a_coding_comparison_query.htm#MiniTOCBookMark10. Published 2021. Accessed October 20, 2021.
24. Wong L. Data analysis in qualitative research: a brief guide to using NVivo. *Malays Fam Physician.* 2008;3(1):14–20.
25. Dollah S, Abduh A, Rosmaladewi R. Benefits and drawbacks of NVivo QSR application. In: Dirawan GD, ed. *Proceedings of the 2nd International Conference on Education, Science, and Technology (ICEST 2017)*. (Advances in Social Science, Education and Humanities Research, vol. 149). Amsterdam, the Netherlands: Atlantis Press; 2017:61–63.
26. Silver C, Bulloch S. CAQDAS at a crossroads: affordances of technology in an online environment. In: Fielding N, Lee R, Blank G, eds. *The SAGE Handbook of Online Research Methods*. London, United Kingdom: SAGE Publications Ltd.; 2017:474–475.
27. Britten N, Fisher B. Qualitative research and general practice [editorial]. *Br J Gen Pract.* 1993;43(372):270–271.
28. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Methods.* 2020;19:1609406919899220.
29. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85(3):257–268.
30. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision: DSM-5-TR*. Washington, DC: American Psychiatric Association; 2022.